UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS

ECONÓMICAS Y EMPRESARIALES

GRADO EN *ECONOMÍA*

TRABAJO DE FIN DE GRADO

ANÁLISIS DESCRIPTIVO Y PREDICTIVO DEL USO DE LA RED DE METRO DE MADRID

AUTOR: *David Villalba Pérez**

TUTOR/ES: *Francisco Álvarez González*

CURSO ACADÉMICO: *2017/2018*

CONVOCATORIA: *Junio*

**email: dvilla01@ucm.es*

Me gustaría acordarme de aquellas personas que han contribuido tanto a este proyecto como a mi época universitaria.

En primer lugar, a mi tutor Paco por guiarme en este proyecto y ser uno de los mejores profesores que he tenido,

A mi equipo de Analytics de Experian, en especial Isa e Iván, por ayudarme y aconsejarme en todo aquello que les he preguntado,

A Rene, por acompañarme, apoyarme y llevarme por el camino correcto en estos años universitarios,

A mi familia entera,

Mi madre por ser el artífice de este proyecto: por la idea que surgió cuando apenas tenía uso de razón de llevarla en Metro hasta la puerta de su trabajo. Este proyecto es por y para ella,

Mi padre por confiar en mi incluso cuando ni siquiera yo lo hacía y ser un apoyo continuo además de mi referente,

Y en especial a mi hermano, por ser el espejo en el que mirarme y que en parte este trabajo es también suyo.

A todos ellos… ¡Muchas Gracias!

ÍNDICE

RESUMEN

Este trabajo utiliza modelos econométricos para explicar las diferencias de uso (número de viajeros) entre las diferentes estaciones de la red de Metro de Madrid. Las variables explicativas consideran tanto factores socioeconómicos (renta y pirámide de edad) del área geográfica donde se sitúa la estación como características de la estación respecto a la red (posición en la línea o existencia de transbordo). Las técnicas econométricas empleadas son arboles de regresión, que permiten segmentar el impacto marginal de cambios en las variables explicativas. Los resultados muestran diferentes patrones de uso entre estaciones dentro la almendra central (interior de la M30) y estaciones de la periferia. Además, se realizan ejercicios de predicción de uso por implantar nuevas estaciones en la red. El modelo considera el impacto que dichas estaciones nuevas tendría en toda la red.

1. INTRODUCCIÓN

El uso del transporte público ha cobrado importancia en los últimos años debido a cuestiones medioambientales y de necesidades de la población.

Este trabajo analiza desde un punto de vista estadístico los determinantes del uso de Metro en Madrid con datos anuales considerando como posibles factores explicativos variables de renta, población y localización geográfica.

La técnica utilizada para realizar el análisis predictivo ha sido, esencialmente, el uso de árboles de regresión. En la realización de la estimación de estaciones extramuestrales se calculó una aproximación del uso de Metro que se haría en aquellas localizaciones geográficas donde actualmente no hay ninguna parada.

Este proyecto queda enmarcado dentro de la literatura relacionada con la economía del transporte, en concreto, en la estimación potencial de la demanda (ver Daniel-Cardozo et al, (2008)). Dentro del sector de transporte, el comportamiento que tiene la demanda con respecto a la utilización que hace del uso de Metro ha sido la principal búsqueda de información realizada. El análisis de la elasticidad en este sector, así como el descenso de la demanda que se ha producido en los últimos años, han sido importantes en tanto que podían ser explicadas por factores inherentes a Metro, o por factores externos.

La demanda del uso de Metro creció de forma constante en el período 1995-2007 (ver Servicio de Planificación y Estudios de Operación (2018)), pero a partir de ese año se produjo un acusado descenso del número de usuarios hasta el año 2015. Dado que no ha habido un cambio significativo en el comportamiento poblacional respecto a la utilización del transporte público, ni tampoco el nivel de uso está muy alejado de la media de las principales ciudades europeas (ver Ayuntamiento de Madrid (2014)), la posible explicación puede venir por la partida destinada por parte del Gobierno Central a los transportes (ver Anibarro García (2016)), así como por la acuciante Crisis Económica que sufre España, en especial en los años en los que los usuarios de Metro tuvieron el mencionado descenso.

En la realización de este proyecto, las técnicas utilizadas para la obtención de datos relacionados con variables Poblacionales y de Renta, fueron: la herramienta Alteryx [1] y el método de asociar dichos datos a las Secciones Censales de la ciudad de Madrid (ver Daniel-Cardozo et al, (2008)). En este último sentido, proyectos similares enmarcados dentro del ámbito de la economía del transporte utilizan también métodos distintos de asignación geográfica de la población, tanto para un posterior análisis de regresión Geográfica (ver Daniel-Cardozo et al, (2012)), como para un análisis múltiple (ver Daniel-Cardozo et al, (2010)). La particularidad de este proyecto es la inclusión de técnicas de Consolidación de datos con Alteryx, el análisis de regresión con lenguaje de programación R [2], así como la utilización de árboles regresivos para obtener un modelo predictivo ajustado para realizar estimaciones de nuevas estaciones extramuestrales.

Las conclusiones obtenidas tras la finalización del proyecto aquí presentado se refieren al comportamiento de la población para usar el Metro y a la predicción de nuevas estaciones. Refiriéndonos al análisis descriptivo de las variables de comportamiento, se puede observar como principal conclusión que los factores para el uso de Metro en estaciones de la Almendra Central y de la Periferia son distintos: mientras que en la Almendra Central prima la conectividad que la estación posee (correspondencia con otras líneas de Metro y transbordo externo, autobuses y Cercanías), en la Periferia debemos incluir, aparte de la conectividad, la cantidad de Población Joven que habita en la zona de influencia de la estación. La predicción concluye que habría un uso positivo en prácticamente la totalidad de las estaciones adicionales extramuestrales tras realizar el estudio de la localización de dichas nuevas paradas de Metro.

El resto del trabajo se estructura de la siguiente forma: En la sección 2 se analiza la Elaboración de los Datos utilizados en el transcurso del proyecto, tanto los procesos de obtención de dichos datos, como el análisis realizado de la demanda de usuarios para cada línea de Metro, o la asignación del total de los datos en distintos Conjuntos para analizar la relación existente entre los factores que explican el Uso de Metro. La sección 3 tiene como fin mostrar el análisis de regresión realizado, con los resultados de cada Conjunto de datos en la sección 4. La comparativa de los resultados se realiza en la

sección 5 para luego ser capaces de afirmar las conclusiones correspondientes dada toda la información obtenida en la realización de este proyecto.

2. ELABORACIÓN DE LOS DATOS

En la obtención de los datos utilizados durante la realización de este proyecto, la principal herramienta que permitió elaborar y unir dichos datos fue el software Alteryx, la cual viene soportada bajo lenguaje de programación R

Se procede a realizar el análisis de los datos utilizados durante este trabajo. Conviene adelantar que el conjunto de datos de los que se alimenta el proyecto los podemos dividir en dos ficheros: por una parte, están los datos asociados a las variables utilizadas para la realización de los modelos regresivos, y por otra parte tenemos el fichero donde se evalúa la distribución que sigue el número de usuarios de Metro en cada estación. Para realizar esta primera parte del proyecto se utilizaron principalmente dos herramientas: Excel y Alteryx. Con Excel se agruparon datos relacionados con el número de usuarios y se crearon gráficas para visualizarlos, y con Alteryx se realizó el proceso de unión de ficheros y consolidación de los datos agrupados en variables. La estructura de comandos que se utilizó para elaborar la elaboración y modificación de los datos en Alteryx se encuentra en el correspondiente Anexo (Alteryx VII) en el final de este fichero.

2.1. Obtención de las variables

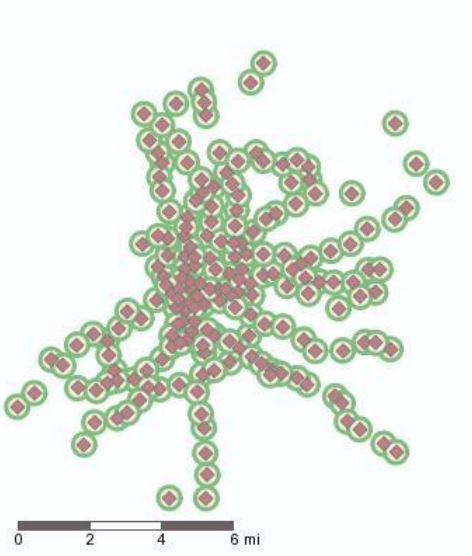
El procedimiento para conseguir las distintas variables no se realizó de forma homogénea para todas las mismas. Hay que distinguir entre las variables relacionadas con la población, que son renta, cantidad de gente por tramo de edad y género, hogares; y las relacionadas con el Metro y las características de cada estación. Para el primer grupo se utilizaron coordenadas geográficas y procesos de consolidación de datos, y para el segundo se procedió a analizar variables potenciales que describieran a las estaciones.

2.1.1. Índices Poblacionales y de Renta

El primer paso consistió en geolocalizar mediante coordenadas cartográficas la situación de cada estación de Metro del municipio de Madrid. Esto consistió en ir

seleccionándolas una a una en *Google Maps* y almacenando la latitud y longitud de todas ellas en un fichero Excel que nos permita incorporarlas en la consolidación de los datos.

Con el objetivo de unir futuras variables numéricas a dicho fichero, debemos convertirlo a un objeto legible para que el sistema lo reconozca: un dato espacial numérico. Con esto obtenemos los puntos de coordenadas puestas en un mapa vacío que llamaré a partir de ahora centroides.

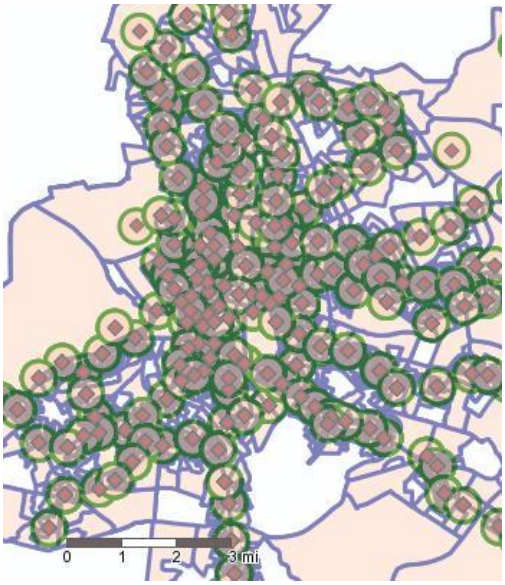


[Figura 1] Centroides con coordenadas cartográficas

Antes de superponer dichos centroides en un mapa cartográfico del municipio de Madrid [3], queda definir el radio de acción de dichos puntos. Yo lo fijé en 0.5 kilómetros de radio, es decir, las estaciones tienen un radio de acción de 500 metros cada una. Es la medida que considero que tienen para influir en la elección de la gente de utilizar dicha parada sin entrar en conflicto con otras que estén próximas.

Definidos los centroides y su radio de acción, llega el momento de cruzar dichos datos con un mapa que contenga información que sea común a los de los demás ficheros. Estos datos comunes que nos van a permitir consolidar todo lo que necesitamos en un solo fichero van a ser las secciones censales, que son unidades territoriales que dividen a la población y organizar los procesos electorales. De esta forma se obtiene un mapa físico con los puntos de los centroides superpuestos en el mapa. Toca ahora dar una medida de la superficie de dichos centroides, en mi caso la medida será en metros cuadrados.

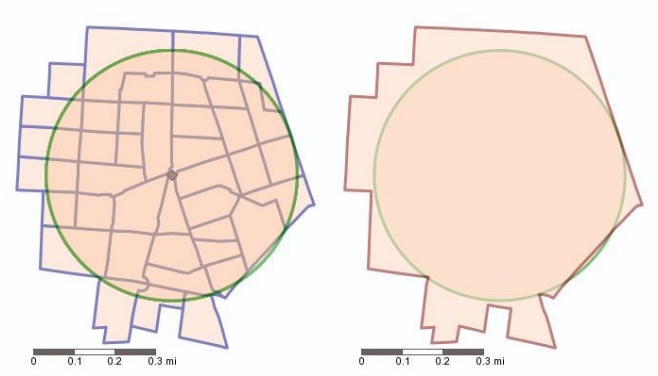
En dicho mapa, el nivel de granularidad alcanzado con las secciones censales [4] nos permite tener información estadística de la población de forma más detallada (ver Cristóbal Pinto et al, (2000)) y comprobar el número de secciones contenidas dentro de nuestros centroides.



[Figura 2] Mapa de secciones censales con los centroides superpuestos

Para poner en orden toda la información, vamos a definir lo que tenemos en este punto. Disponemos de tres capas distintas, que son: área de la sección censal entera, área del Centroide, y área que conforman la intersección entre el Centroide y la sección censal.

En el proceso previo a la consolidación final, hay que definir qué secciones censales deben estar incluidas en el centroide estándar. En mi caso, he decidido que se contabilizará como influyente dentro del centroide aquella sección censal que ocupe el 30% del Centroide, o bien si el área de la sección censal que está contenido en el Centroide es mayor de 5% del total de la sección (esto lo definí así para que todas las estaciones tuvieran un rango de acción y no dejar a ninguna sin datos). El paso final es hacer el sumatorio de todas esas secciones censales influyentes en los centroides y obtener el conglomerado de acción final con lo toda la información relativa a la población.



[Figura 3] Ilustración de centroide dividido en secciones censales (izquierda) y conglomerado total (derecha)

Obtenidos los polígonos con la información de cada estación, vamos a adherirles los datos correspondientes a Población y Renta. El método seguido fue crear dos nuevos ficheros de datos: uno que contenga datos de renta y el otro de Población. El fichero con los datos de renta lo obtuve a través de una base de datos [3] con fecha más actualizada del año 2014, en ese fichero se encuentran los datos de renta per cápita, renta de los hogares, y la enumeración de las secciones censales de la ciudad. El fichero con los datos Poblacionales contiene datos de número de hogares, cantidad de población por tramos de edad, género de cada uno de los individuos y población total nacional y extranjera, además de las secciones censales enumeradas como en el caso anterior. Estos datos poblacionales han sido obtenidos buscando y añadiendo información a dicho fichero a través de varias bases de datos abiertas [4] [5] [6] y datan de 2016, son las más actualizadas encontradas y provienen del censo municipal.

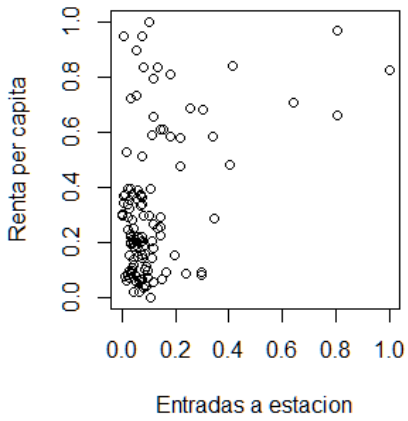
A partir de los datos de secciones censales que tiene cada fichero, se procede a la unión final de la información. De este modo tenemos los índices de **Población Total**, **Renta per Cápita**, **Población Joven** (de 10 a 24 años), **Población Adulta** (de 25 a 65 años), **Población Sénior** (de 66 a 85 años), **Número de Hogares**, **Renta media por Hogar**, **Población total de varones**, **Población total de mujeres**, **Población Nacional** y **Población Extranjera**.

Aunque no usaremos todas las variables obtenidas, más adelante se mostrará otra función que se le ha dado a estos índices.

2.1.2. Índice de Uso de Metro

Esta variable es la utilizada como variable dependiente de la regresión. Se analizará su variación según los factores escogidos.

El proceso de obtención fue a través de una base de datos abierto [7] donde pude encontrar el número de usuarios que entraron en cada estación de la Red de Metro de Madrid. Estos datos son una foto fija de la fecha de estudio: 2017, y ya sólo con ellos podemos observar comportamientos en las entradas a las estaciones.



[Figura 4] Relación entre Uso de Metro (Entradas a la estación) y la renta per cápita de la zona de influencia. Los usuarios que usan el Metro no tienen una renta disponible alta.

2.1.3. Índices Relacionados con el Metro

El siguiente grupo de índices es el relacionado con las características de las estaciones [8]. Este conjunto de datos podemos dividirlo en dos partes:

- Variables ficticias o *Dummies*: estas variables tienen carácter dicotómico y se crearon en base a los datos obtenidos de las características del metro. Dichos datos corresponden a **Transbordo de Metro en la Estación** (valor 1 si lo hay y 0 en caso contrario); **Transbordo con Red de Transporte Externa**, que puede ser desde Cercanías, Metro Ligero, Metro Norte, Metro Oeste, trenes de larga distancia o regionales, o intercambiador de autobuses interurbanos; **Estación Adaptada a Minusválidos**, este valor toma 1 si lo es, y 0 si no lo es, y puede ser una variable proxy con una estación nueva o reformada y modernizada; **Estación**

con Párking, valor 1 si cuenta con ello y 0 si no; y **Zona de Influencia Turístico-Cultural-Comercial** [9], si la estación está situada en una zona con importante valor de este tipo (1) o si no (0).

- Variable Real: únicamente encontramos como real la variable **Número de Estaciones Restantes hasta Transbordo**, que contabiliza el número de paradas que faltan hasta encontrar una que tenga correspondencia con otra estación perteneciente a la Red de Metro.

2.1.4. Normalización de las variables

Para poder leer la totalidad de datos de una forma más eficiente, se realizó el proceso de normalización. Dados los valores que tienen determinadas variables, como puede ser las entradas a una estación (valores en millones de entradas), o los datos de población (en miles), en contraste con las variables Dummy o los ratios, se consideró conveniente realizar dicha conversión en score para homogeneizar resultados.

La fórmula utilizada para normalizar es la siguiente:

$$X = \frac{\gamma_i - \gamma_{min}}{\gamma_{max} - \gamma_{min}} \quad (1)$$

siendo X el valor normalizado, γ_i la observación del parámetro con su máximo y su mínimo correspondientes a todas las observaciones de dicha variable.

2.1.5. Ratios obtenidos con relaciones entre variables

A partir de los datos obtenidos con el proceso de elaboración de variables detallado en anteriores apartados, se procedió a crear nuevas variables que pudieran ser susceptibles de tener significancia respecto al uso de Metro. Estos índices se obtienen relacionando variables con potencial sentido con respecto a la población.

- **Ratio de Zona en Potencial Auge**: Marca el grado de crecimiento conjunto que puede tener una zona en estudio, la zona de influencia de la estación. Tomará valores altos (cercanos a 1) para aquellas zonas con un mayor índice de expansión. Esta variable ha sido creada como un ratio entre:

$$\frac{Población\ Joven + Población\ Adulta}{PoblaciónTotal(Joven + Adulta + Sénior)} \quad (2)$$

- **Ratio Potencial de Tasa de Población en Edad de Trabajar:** Esta tasa puede ser una aproximación de la tasa de actividad, aunque dicha fórmula contempla una edad para trabajar menor a la que aquí está contabilizada:

$$\frac{Población\ Adulta}{PoblaciónTotal(Joven + Adulta + sénior)} \quad (3)$$

- **Ratio de Envejecimiento de la Población**: Este ratio indica el nivel de envejecimiento de la zona de influencia.

$$\frac{Población\ Sénior}{Población\ Joven} \quad (4)$$

- **Densidad de Habitantes por Hogar**: Valor que marca el número de habitantes por hogar dentro de la zona de influencia.

$$\frac{Población\ Total}{Número\ de\ Hogares} \quad (5)$$

- **Ratio de Género de la Población:** Marca la incidencia de la población de género femenino en la zona estudiada.

$$\frac{Población\ Total\ Mujeres}{Población\ Total\ Hombres} \quad (6)$$

- **Ratio de Inmigración:** Indica el nivel de inmigración existente en la zona de influencia.

$$\frac{Población\ Nacional}{Población\ Extranjera} \quad (7)$$

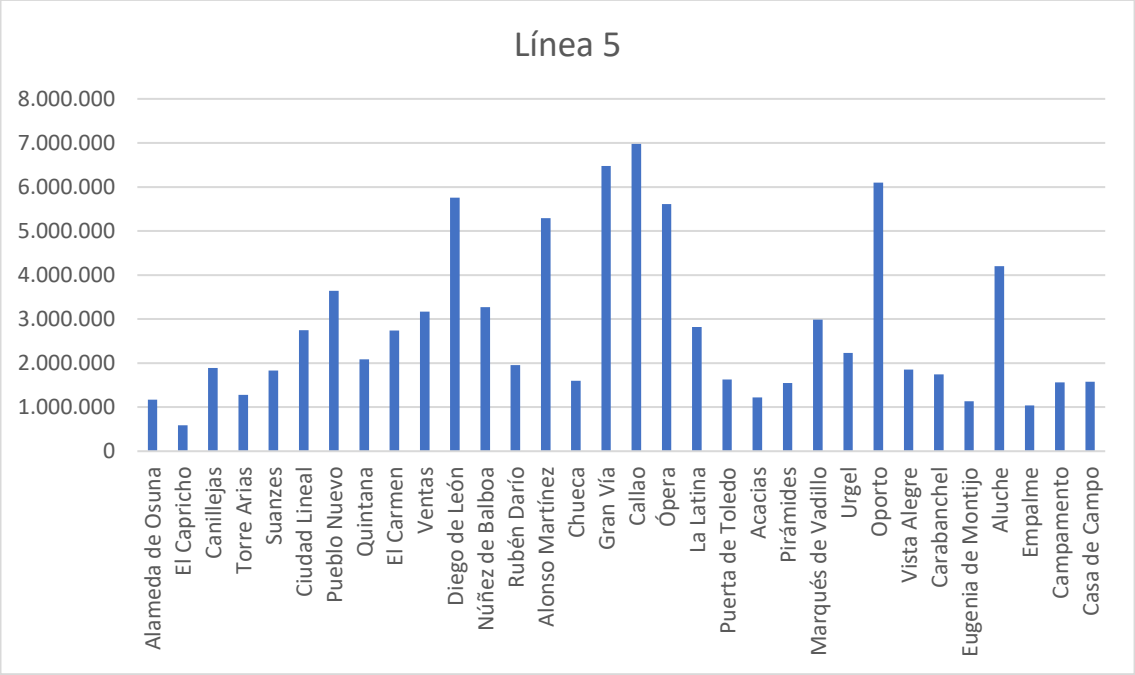
En la tabla mostrada en el Anexo I podemos observar las variables que he utilizado en el desarrollo del proyecto, ya que no todas las variables que han sido generadas se han aplicado.

2.2. Distribución de usuarios de Metro de Madrid

La segunda parte del análisis de los datos se centra en las distribuciones de usuarios de las líneas de metro [10].

El estudio realizado incluye las distribuciones de usuarios en cada estación por trayecto de las 10 líneas de la Red de Metro de Madrid (se excluye la 6, Circular). El fin de este análisis es comprobar la normalidad en las distintas distribuciones, además de cuantificar el número de usuarios que utilizan estaciones en los extremos del trayecto de cada línea. Para ello se ha establecido que el **30%** de la distribución de las estaciones serán los extremos, y el **40%** la parte central, siguiendo el número total de estaciones de la línea. Si la distribución tuviera un alto grado de normalidad, teniendo en cuenta que las estaciones con un mayor volumen son las que se sitúan en la parte central, y las que tienen un volumen menor de usuarios son las que deberían estar tanto en el extremo derecho como en el izquierdo, habría un gran aprovechamiento de la línea, que no acepta una estación adicional en el extremo.

La regla que he utilizado para definir extremo izquierdo o extremo derecho ha sido la localización geográfica de la trayectoria: el extremo izquierdo son trayectorias en la que la cabecera de línea está en el Norte, y el derecho son aquellas cuya cabecera está en el Sur.



[Figura 5] Distribución de usuarios de la línea 5 de Metro

En la Figura se puede observar un ejemplo de línea que **no** admitiría nuevas estaciones dado que su distribución tiene una alta Normalidad., los extremos no tienen

alta afluencia de pasajeros. Mientras que en el Anexo II observamos un ejemplo de línea que sí podría admitir una nueva estación en su extremo derecho.

2.3. Conjuntos de Datos

Si se quiere captar y comparar la variabilidad en las entradas a la estación, es necesario ver los resultados obtenidos con varios Conjuntos de datos distintos. En base a los resultados obtenidos, se elegirá el modelo que mejor explique el valor del Uso de Metro.

En este caso he creado tres Conjuntos de datos para analizar más adelante en el proyecto su incidencia y los resultados comparándolos entre ellos.

2.3.1. Conjunto 1: Total de los Datos

El primer Conjunto analiza el total de las observaciones. No obstante, tras analizar la muestra, se comprobó la existencia de datos atípicos. La acción realizada fue eliminar estas cuatro estaciones que casi no tenían cobertura a nivel de radio de acción, y por tanto aportaban poca información relativa a población o renta. Estas estaciones fueron Ciudad Universitaria, Aeropuerto T1-T2-T3, Aeropuerto T4 y Feria de Madrid. En adición, al realizar las pruebas de Homocedasticidad con las regresiones, se comprobó que tanto el problema de Heterocedasticidad como el de No Normalidad del Error [11] (ver Sotoca (2012)) mejoraban con la ausencia de estas cuatro estaciones. Así pues, nos quedamos con un conjunto de datos compuesto por 188 observaciones.

2.3.2. Conjunto 2: Estaciones de la Periferia

Este segundo conjunto de datos analiza la muestra a partir de estaciones que están situadas fuera de la almendra central (fuera de la M-30). Comparándolo con el anterior, el número de estaciones a analizar se reduce ostensiblemente, pero el resultado de la importancia de las variables en el modelo se intuye que va a ser diferente. Se tienen ahora 99 observaciones.

[Figura 6] Mapa de la Almendra Central (M-30) [12]

2.3.3. Conjunto 3: Datos de las estaciones en los extremos de líneas

En este tercer conjunto de datos vamos a tomar como observaciones aquellas estaciones situadas en ambos extremos de las líneas, un total de 115 estaciones a evaluar.

3. ANÁLISIS DE REGRESIÓN

Se procede a realizar las pruebas pertinentes de comprobación de la variabilidad del Uso de Metro para cada uno de los Conjuntos de datos vistos con anterioridad.

El objetivo principal será comprobar si las variables influyen de igual modo para todos los conjuntos de datos o si, por el contrario, el hecho de tener estaciones situadas en distintos puntos arroja resultados dispares. Dada la ausencia de temporalidad en los datos, ya que contamos con una foto fija tanto de usuarios de Metro (2017), como Renta (2014) y Población (2016), el modelo usado es el Lineal Regresivo Múltiple [11] (ver Sotoca (2012)) para comprobar tanto el nivel descriptivo como el predictivo de las variables independientes.

En este apartado se muestra la Metodología con la que se realizaron los análisis de regresión con los que se trabaja a lo largo del proyecto.

Esta parte del proyecto fue realizada con el software de programación libre R, tanto para manipular datos como para ver la influencia que éstos tenían en los distintos conjuntos. Las librerías utilizadas durante el transcurso del trabajo están documentadas en la sección bibliográfica final (ver Grumping (2006), Hofner (2017), Kuhn (2018), Peters y Hothorn (2017), Revelle (2018), Wei y Simko (2017), Wickham (2009), Wickham (2017), Wickham et al. (2017), Xie, Zeileis (2004), Zeileis y Hothorn (2002)).

El método utilizado es homogéneo para todos los modelos. Se parte de un análisis de los datos (ver Guisande González y Vaamonde Liste (2012)), a continuación, se comprueba la influencia que tiene cada variable en el modelo regresivo mediante el análisis descriptivo (ver Méndez Suárez (2018)), para finalizar con el nivel predictivo que tiene cada uno de los conjuntos de datos.

3.1. Análisis de los Datos

Después de cargar en nuestro software el Data Set (conjunto de datos) pertinente se realizó un análisis de los datos que manejamos.

Se comprobaron las características de las variables cargadas y se analizó la existencia de algún *missing* o dato no válido en ellas.

Con el fin de evaluar el comportamiento de las variables, y dada la linealidad en los datos, el estudio de la correlación se realizó con el Coeficiente de Correlación de Pearson. Con la Matriz de Correlaciones se trató de agrupar las variables con mayor índice relativo. Pero que pertenezcan al mismo grupo y exista correlación no es necesariamente algo negativo, lo que hay que analizar es la existencia de la colinealidad entre factores.

Para poder localizar mejor aquellos grupos que contenían variables correlativas se realizó el test de Correlación de Pearson individual entre aquellas variables potenciales por separado, lo que permite tratar con mayor cuidado dichas variables en pasos posteriores y estar alerta ante posibles casos que vicien el modelo.

3.2. Análisis Descriptivo

En la parte del Análisis Descriptivo (ver Méndez Suárez (2018)), lo que se hizo en primer lugar fue crear en R la regresión lineal que engloba nuestras variables: la explicada (Uso de Metro), y el conjunto de variables dependientes (explicativas) obtenido con anterioridad.

Se procedió a la comprobación de la existencia de la multicolinealidad mediante la fórmula de *Variance Influence Factors* (VIF), que consiste en el cálculo de la correlación muestral entre variables explicativas. Lo que nos va a mostrar este valor va a ser si tenemos nula o leve colinealidad (VIF entre 1 y 5), colinealidad moderada (VIF entre 5 y 10), o VIF alta (valores mayores que 10).

$$VIF_j = \frac{1}{1 - {R_j}^2} \tag{8}$$

Siendo VIF_j el valor para cada variable, y ${R_j}^2$ el coeficiente de regresión del j-ésimo regresor sobre el resto [11] (ver Sotoca (2012)).

En los datos se comprobó la existencia de multicolinealidad para variables de tipo poblacional entre los datos de Población adulta, Sénior, Número de Hogares, y para las variables de Renta: renta per cápita y Renta por Hogar. Este proceso de mejorar la calidad de los datos consistió en eliminar de forma sistemática aquellas variables en las que VIF detectaba multicolinealidad y se realizó la consiguiente regresión con las variables restantes. En el momento en el que nuestras variables tenían un nivel de correlación que no entrañaba posible multicolinealidad, pasaríamos a comprobar la significación de las variables en la regresión.

Con el fin de establecer un diagnóstico fiable de la regresión, se deben seleccionar las variables que tienen un grado de significancia aceptable. El nivel de significación para las variables lo he marcado en un p-valor del 15%. Esto quiere decir que únicamente se van a aceptar como variables significativas para el modelo aquellas que tengan un p-valor menor de 0.15. La fijación de este umbral de significancia se debió a la poca cantidad de observaciones con las que se contaba y si se marca un valor crítico menor corremos el riesgo de perder información.

El método utilizado para realizar el diagnóstico de significancia de la regresión consistió en tres procesos: primero se elimina aquella variable con el mayor p-valor de cuantas variables contaba, luego se realiza la nueva regresión con las variables restantes, y por último utilizamos dos test para probar si la nueva regresión es más completa que la anterior. Para este último paso se utilizaron tanto el criterio de información de Akaike, el cual nos indica, mediante una disminución de su valor, si el modelo acepta el cambio de variables realizado.

$$AIC(k) = log[{\sigma_k}^2] + \frac{2\,k}{n} \tag{9}$$

Siendo AIC el criterio de información de Akaike, ${\sigma_k}^2$ la estimación de la varianza residual por máxima verosimilitud, k el número de variables independientes del modelo, y n el tamaño de la muestra.

También se realizó el test de ANOVA (*analysis of variance*), el cual nos indica mediante una distribución F de Snedecor los efectos en la varianza que vamos a tener al eliminar una variable. El test de ANOVA se trata de realizar un contraste de significación donde la Hipótesis Nula será la aceptación de que las Medias Poblacionales son iguales, frente a la Alternativa donde al menos encontramos dos Medias distintas entre sí. Un p-valor alto frente a la distribución F determina que aceptamos la hipótesis de igualdad de varianzas, y por tanto no vamos a tener cambios significativos en el modelo al eliminar la variable sobrante.

Este paso se repite tantas veces como variables no significativas tengamos, hasta obtener un modelo ajustado completamente. Esto se puede acelerar en R al poner el comando “trace=0”, que nos evita los pasos de modelos intermedios hasta obtener el final, además de ser el que mayor valor de R cuadrado tiene. Esto es así debido a que se consigue explicar la variable dependiente al máximo con las variables que han quedado.

Una vez diagnosticadas las variables del modelo, se procede a evaluar la situación de los residuos que tenemos. Para ello se comprueba el nivel de Homocedasticidad, la varianza de los residuos que tiene el modelo regresivo. Un nivel constante en la varianza de los residuos a lo largo de las observaciones denotaría Homocedasticidad, en cambio si no es constante tendríamos Heterocedasticidad.

La forma que he utilizado para comprobarlo es a través del test de Breusch-Pagan, en el cual se contrasta bajo Hipótesis nula de Homocedasticidad. Si el p-valor resultante es mayor de 0.05 podemos afirmar la Nula de Homocedasticidad en el modelo. En caso de no llegar a esta conclusión, lo que hice fue analizar la gravedad de la Heterocedasticidad con el Contraste de White para comprobar el nivel de robustez. Para ello realizamos la matriz de varianzas y covarianzas y evaluamos los valores de la diagonal de la matriz. Realizamos la raíz cuadrada de los mismos y dividimos los coeficientes entre éstos. El resultado será un valor distribuido en una t de Student que interpretaremos como medida de robustez: cuanto más cercanos a cero sean los valores obtenidos, menos robustas serán las variables correspondientes, lo que nos llevaría a eliminar del modelo las variables que sean más críticas.

Una vez obtenida la matriz de varianzas y covarianzas:

$$WhiteTestValues_i = \sqrt{diag[Var - Cov\ matrix]} \tag{10}$$

$$\frac{\beta_i}{WhiteTestValues_i} \sim t_{(n-k)} \tag{11}$$

Siendo β_i los coeficientes que tenemos en nuestra regresión ajustada y $WhiteTestValues_i$ los valores obtenidos de hacer la raíz cuadrada de la diagonal de la matriz de varianzas y covarianzas.

Por último, se realiza un Histograma representativo de los residuos que tenemos para comprobar la normalidad en los mismos.

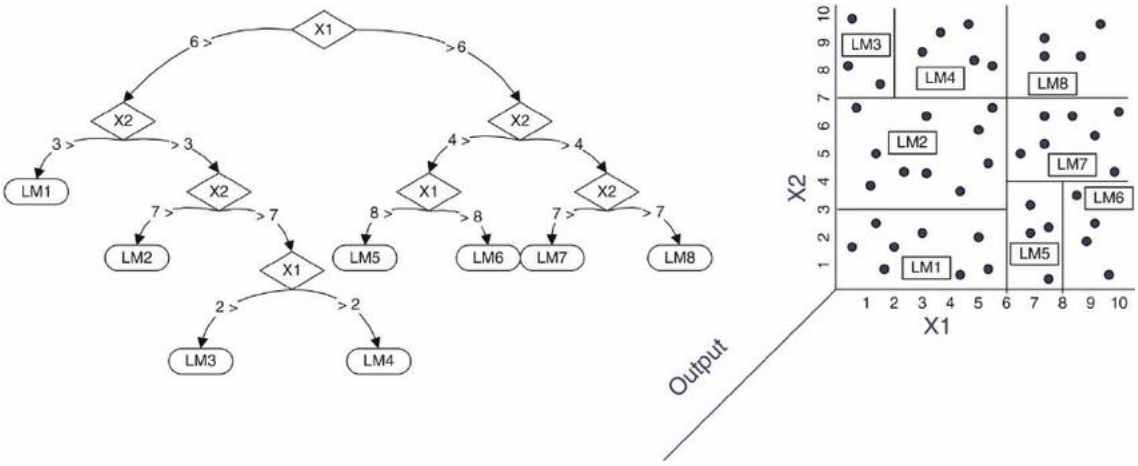
Una vez comprobados estos parámetros de los residuos, se procede a visualizar la importancia relativa que tienen las variables independientes sobre la explicada, y en qué grado porcentual influyen en la misma.

3.3. Análisis Predictivo

Analizar la importancia relativa de los factores sirve de base para la última parte del análisis de regresión, aunque dependiendo del conjunto de datos que utilicemos hará que se juste más a unas variables que a otras. El método utilizado es el de "auto entrenamiento" de nuestro modelo.

Lo primero que hay que hacer es dividir nuestro Data Set y fijar el porcentaje que vamos a destinar para entrenamiento, y para test final. En mi caso he fijado en un 83% del Data Set para que el modelo pueda entrenarse, y el 17% restante para testar el ajuste del modelo.

La forma en la cual se va a entrenar el modelo es a través de un árbol de regresiones. El método elegido ha sido el llamado M5P (ver Lantz (2015)), que es un árbol de nodos de decisión para modelos lineales. Este método discrimina entre los nodos más altos para ir llegando a las "ramas más bajas". El proceso de ajuste trata de definir la recta de regresión en distintos tramos según el número de regresiones auxiliares que el modelo determine necesario.



[Figura 7] Ejemplo de árbol de regresiones [13]

A partir de la "copa del árbol" empieza a discriminar por las ramas según el resultado que arroje el valor de la primera regresión. La característica principal del modelo M5P es que se trata de un árbol de regresiones: al realizar la primera regresión, el resultado se filtra a través del nodo hasta llegar a las diversas ramas, donde se vuelven a hacer nuevas regresiones auxiliares con coeficientes distintos según la discriminación que haga el nodo de decisión (X1, X2) y se repite el proceso (LM1, LM2, etc.) hasta obtener el resultado del modelo predictivo final.

Obtenido el modelo a partir del árbol, el proceso fue introducir en una función predictiva dicho modelo junto con el conjunto de datos reservado para testarlo. Esta función predictiva va a devolver el valor de la variable dependiente correspondiente para cada observación.

El análisis de la capacidad de predicción de los distintos modelos se realizó con el Error Cuadrático Medio y su correspondiente visualización. Consiste en una comparación entre los valores predichos que nos devuelve el modelo y los reales, para obtener una media del error en base a las observaciones. Cuanto más pequeño sea el Error, más exacto será el modelo en cuanto a capacidad predictiva se refiere.

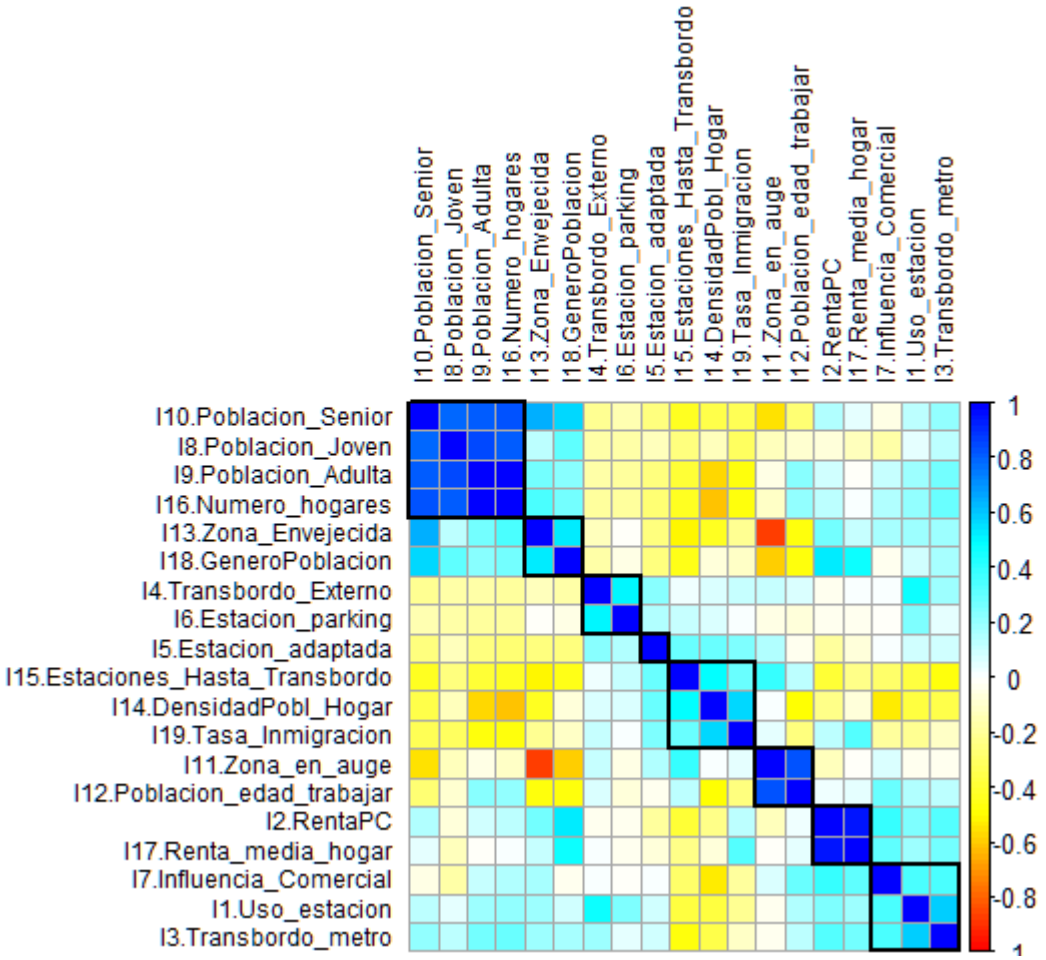
4. RESULTADOS POR CONJUNTOS

4.1. Conjunto 1: Datos con el total de las observaciones sin datos atípicos.

4.1.1. Análisis de los Datos

Tras cargar los datos correspondientes al Modelo 1, efectuamos la matriz de correlaciones poniendo especial hincapié en aquellas variables con un alto grado de correlación individual entre ellas.

En este caso, lo que se puede observar es un alto grado de correlación entre los grupos de Renta: Renta per Cápita y Renta media del hogar, y el grupo Poblacional: Poblaciones por edades, número de hogares y ratios de población. También se observa correlación lógica entre la Zona en Auge y la población en edad de Trabajar.



[Figura 8] Matriz de Correlaciones del Conjunto 1

4.1.2. Análisis Descriptivo

Fijadas aquellas variables susceptibles, procedemos a analizar descriptivamente este modelo. Lo que se deduce tras aplicar la fórmula de VIF (Factor de inflación de la varianza), es que tenemos multicolinealidad en nuestra regresión. Quitando una a una las variables con un mayor valor de VIF (mayor de 8), y tras realizar los contrastes de significación de las variables, obtenemos un modelo ajustado final donde tenemos hasta seis variables independientes que explican el Uso del metro, en un modelo con un R-cuadrado o bondad del ajuste del 55.2%.

COEFFICIENTS	ESTIMATE	STD. ERROR	T-VALUE	Pr(>|t|)
(Intercept)	0.310	0.091	3.408	0.001
I2.RentaPC	0.065	0.029	2.265	0.025
I3.Transbordo_metro	0.135	0.020	6.862	1.05e-10
I4.Transbordo_Externo	0.151	0.019	7.956	1.87e-13
I5.Estacion_adaptada	0.024	0.015	1.552	0.123
I14.DensidadPobl_Hogar	-0.107	0.041	-2.613	0.010
I19.Tasa_Inmigracion	-0.004	0.002	-2.307	0.022

Residual standard error	0.093 on 181 degrees of freedom
Multiple R-squared	0.567
Adjusted R-squared	0.552
F-statistic	39.45 on 6 and 181 Degrees of freedom, p-value: < 2.2e-16

[Figura 9] Regresión ajustada del Conjunto 1

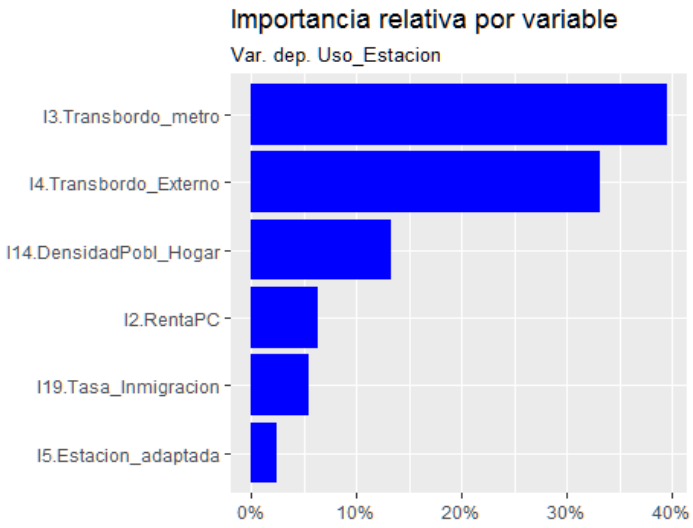
Los resultados son los esperados en el análisis descriptivo de las variables independientes.

Además, la medida de VIF nos arroja que las variables del modelo ajustado tienen un grado leve de correlación con un valor máximo de 1.4, el criterio de Akaike comparativo entre el modelo inicial y el ajustado final es de 6 puntos, y el cálculo del análisis de la varianza resulta que el modelo no se resiente al eliminar las variables no significativas, lo resulta en una mayor capacidad predictiva final (Anexo III [Figura 24]).

Tras realizar el test de Breusch-Pagan no se detecta Homocedasticidad, pero sí que comprobamos que no hace falta eliminar ninguna variable realizando el Contraste de White y obteniendo todos los valores aceptables en la ejecución de la robusta Homocedasticidad, no tenemos ninguna variable crítica en este sentido.

En el análisis de los residuos, como se puede observar en la [Figura 25] (Anexo III), tenemos cierta normalidad en el Histograma a pesar de tener ciertos datos atípicos como arroja dicha gráfica.

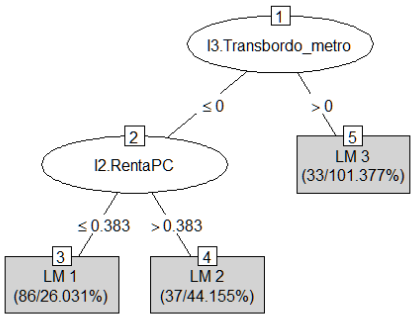
En el cálculo de las variables más importantes dentro de nuestro modelo, cabe destacar que las variables más importantes en este modelo son el grado de correspondencia con red de Metro que tiene la estación, y la existencia de Transbordo con red de transportes externa.



[Figura 10] Importancia relativa de las variables significativas en el Conjunto 1

4.1.3. Análisis Predictivo

Después de realizar la partición correspondiente entre prueba de entrenamiento y testeo de los datos, se realiza el modelo predictivo mediante el método de árboles regresivos M5P explicado anteriormente, y seleccionando las variables que únicamente han resultado ser significativas en la parte Descriptiva, realizamos el modelo de decisión.



[Figura 11] Modelo de árbol de regresiones M5P para el Conjunto 1

Acabamos el entrenamiento de los datos con la predicción de la variable dependiente Uso de Metro y medimos el Error Cuadrático Medio del modelo. Tras realizar las operaciones de normalización inversa resulta que este modelo tiene un error en la predicción de 687399.67 usuarios por estación en el flujo de un año (Anexo III [Figura 25]), respecto a los 2962436.681 usuarios de media por estación en este Conjunto de datos.

4.2. Conjunto 2: Datos correspondientes a estaciones fuera de la almendra central (M-30)

4.2.1. Análisis de los Datos

Se valoran las correlaciones existentes entre nuestras variables en este conjunto de datos para poner especial atención en ellas.

Como se puede observar en la [Figura 26] (Anexo IV), en este caso destaca la correlación existente entre los grupos asociados a renta y el ratio de habitantes nacionales y extranjeros; así como entre la zona de población envejecida y el ratio entre el número de mujeres y hombres en la población de la zona.

4.2.2. Análisis Descriptivo

Tras realizar la prueba de multicolinealidad y eliminar variables que sufrían este problema, y ajustar el modelo eliminando las variables no significativas, nos quedamos con un modelo lineal regresivo con nueve variables independientes y con el intercepto que, aunque su p-valor es mayor de 0.15 no es influyente en el modelo.

COEFFICIENTS	ESTIMATE	STD. ERROR	T-VALUE	Pr(>|t|)
(Intercept)	-0.395	0.296	-1.336	0.185
I3.Transbordo_metro	0.284	0.054	5.200	1.26e-06
I4.Transbordo_Externo	0.171	0.037	4.636	1.21e-05
I5.Estacion_adaptada	-0.043	0.027	-1.582	0.117
I6.Estacion_parking	0.110	0.049	2.263	0.026
I8.Poblacion_Joven	0.338	0.066	5.088	2.00e-06
I12.Poblacion_edad_trabajar	0.702	0.363	1.935	0.056
I13.Zona_Envejecida	0.059	0.040	1.471	0.145
I15.Estaciones_Hasta_Transbordo	-0.011	0.006	-1.831	0.070
I19.Tasa_Inmigracion	-0.004	0.002	-1.878	0.063

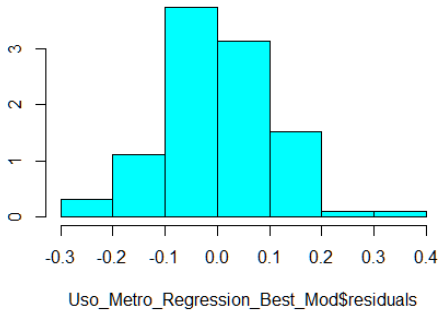
Residual standard error	0.115 on 89 degrees of freedom
Multiple R-squared	0.677
Adjusted R-squared	0.644
F-statistic	20.74 on 9 and 89 Degrees of freedom, p-value: < 2.2e-16

[Figura 12] Regresión ajustada Conjunto 2

El valor del R-cuadrado ajustado es de 64.4%, lo que lo convierte en el modelo más completo en este sentido de cuantos se han estudiado en el proyecto; en la medida de la varianza y el criterio de Akaike (Anexo IV [Figura 27]) se observa una mejora del modelo ajustado respecto al inicial, y en lo que se refiere a multicolinealidad, el valor máximo de 2.8 de VIF en la variable de Zona Envejecida muestra que las variables son independientes unas de otras por norma general.

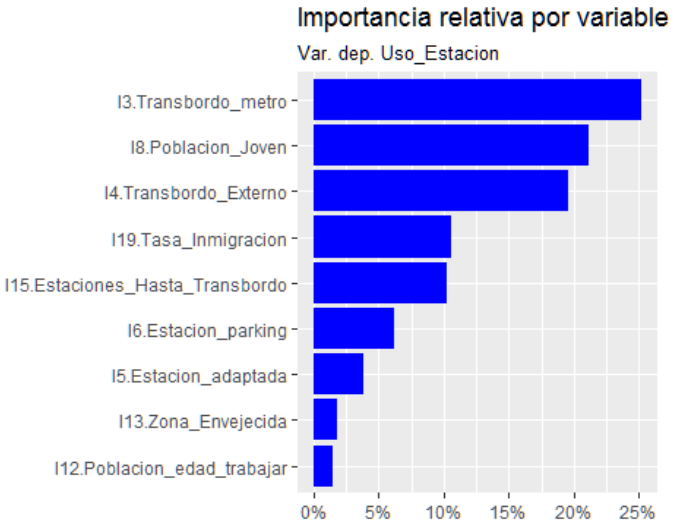
En el análisis de la Homocedasticidad, al realizar el test de Breusch-Pagan se observa que el nivel del p-valor es 0.001 con lo que rechazamos la hipótesis nula de Homocedasticidad, aunque es un valor próximo al límite lo que nos hace pensar que el problema no será grave, así se desprende tras realizar el contraste de White de la robusta Homocedasticidad y ver resultados correctos.

El estudio de los residuos, con el Histograma, observamos nuevamente cierta normalidad.



[Figura 13] Histograma de los Residuos en el Conjunto 2

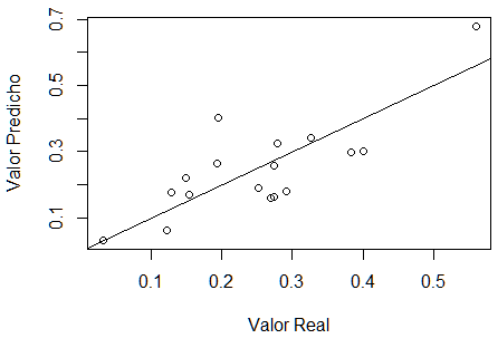
Refiriéndonos a la importancia de las variables dentro del modelo, aquellas con mayor peso son Transbordo de red de metro, Densidad de Población Joven, y transbordo con red externa a metro.



[Figura 14] Importancia relativa de las variables Conjunto 2

4.2.3. Análisis Predictivo

Al realizar el modelo predictivo con este conjunto de datos, comprobamos que el Error Cuadrático Medio es de 0.0079 valor normalizado. Sin normalizar, el modelo actúa con un error medio de 387718.65 usuarios anuales por estación, frente a los 1911924.576 usuarios de media por estación en este Conjunto de datos.



[Figura 15] Gráfica de Error Cuadrático Medio del Conjunto 2

4.3. Conjunto 3: Datos correspondientes a Estaciones en los extremos derecho e izquierdo de la distribución

4.3.1. Análisis de los Datos

Analizando las características de este conjunto de datos en lo que a correlaciones se refiere, es muy parecido al primer modelo (Anexo V, [Figura 28]), donde vemos muy definidas las agrupaciones de variables entre ellas: grupo poblacional, de Renta y ratios de tasa de actividad y zona envejecida.

4.3.2. Análisis Descriptivo

Siguiendo el procedimiento utilizado en los anteriores modelos, nos disponemos a evaluar el grado de multicolinealidad de las variables y su nivel de significación.

En el modelo ajustado contamos con ocho variables explicativas del Uso de Metro, ausencia de multicolinealidad con un VIF máximo de 1.8, un criterio de Akaike (Anexo V [Figura 29]) que mejora 9.6 puntos sobre el modelo inicial, y un nivel de R-cuadrado de 60.7%, capacidad notable de explicar el Uso de Metro por parte de las variables independientes.

COEFFICIENTS	ESTIMATE	STD. ERROR	T-VALUE	Pr(>|t|)
(Intercept)	-0.064	0.054	-1.196	0.235
I2.RentaPC	0.121	0.046	2.652	0.009
I3.Transbordo_metro	0.150	0.030	5.057	1.80e-06
I4.Transbordo_Externo	0.151	0.025	5.966	3.25e-08
I5.Estacion_adaptada	0.035	0.022	1.597	0.113
I7.Influencia_Comercial	0.054	0.030	1.788	0.077
I8.Poblacion_Joven	0.087	0.052	1.691	0.094
I13.Zona_Envejecida	0.037	0.021	1.730	0.090
I19.Tasa_Inmigracion	-0.004	0.002	-2.253	0.026

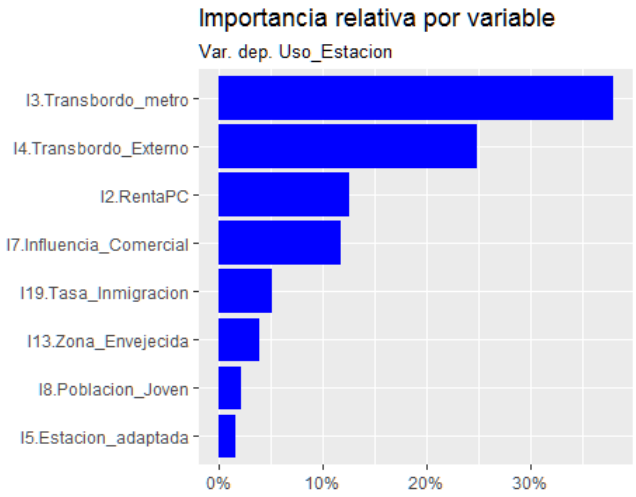
Residual standard error	0.100 on 106 degrees of freedom
Multiple R-squared	0.634
Adjusted R-squared	0.607
F-statistic	22.99 on 8 and 106 Degrees of freedom, p-value: < 2.2e-16

[Figura 16] Regresión del Conjunto 3

Detectamos un problema de no Homocedasticidad al realizar el test de Breusch-Pagan y rechazar la hipótesis nula con un p-valor de $1.94*10^{-6}$, aunque el Contraste de White no nos muestra un problema grave al no mostrarnos ningún valor crítico para ninguna variable, seguimos con las ocho variables significativas.

Es quizás en este modelo donde encontramos el mayor valor de no normalidad en la distribución de los residuos, algo insalvable con los datos elegidos (Anexo V [Figura 30]).

A la hora de analizar la importancia relativa de las variables en el modelo, en este caso el Transbordo con Red de Metro, Transbordo Externo, Renta per cápita y nivel comercial en la zona son las variables más influyentes. Esto puede deberse a la conexión entre las partes alejadas y las más céntricas de la línea de metro. Considerando que las ampliaciones únicamente se pueden dar en los extremos de la línea, un mayor uso de Metro se da en estaciones con correspondencia entre líneas o con transbordo con transporte interurbano.



[Figura 17] Importancia relativa de las variables del Conjunto 3

4.3.3. Análisis Predictivo

Incorporamos las ocho variables significativas a nuestro modelo de árbol de regresiones para obtener el criterio de decisión tras el entrenamiento del modelo. El nodo de decisión que obtenemos del modelo está en la variable Influencia comercial en la zona (Anexo V [Figura 31]), de ahí salen dos regresiones en cada rama del nodo para la elección final.

El error Cuadrático Medio en valor normalizado es de 0.0044, mientras que el valor sin normalizar y adaptado a usuarios totales va a ser de 563066.061 (Anexo V [Figura 30]), frente a los 2714063.087 usuarios de media por estación en este Conjunto de datos.

5. COMPARATIVA ENTRE CONJUNTOS DE DATOS

La comparativa entre los tres conjuntos de datos pone de manifiesto las diferencias existentes entre las distintas zonas donde están situadas las estaciones. Dependiendo de cuál de los tres se escoja va a arrojar un nivel u otro tanto a nivel descriptivo de los factores más importantes como de estimación para obtener una aproximación en el uso de Metro en estaciones nuevas extramuestrales.

5.1. Determinantes de Uso

[Figura 18] Comparativa de las variables más importantes en cada Conjunto. A la izquierda Conjunto 1, en el centro Conjunto 2, a la derecha Conjunto 3

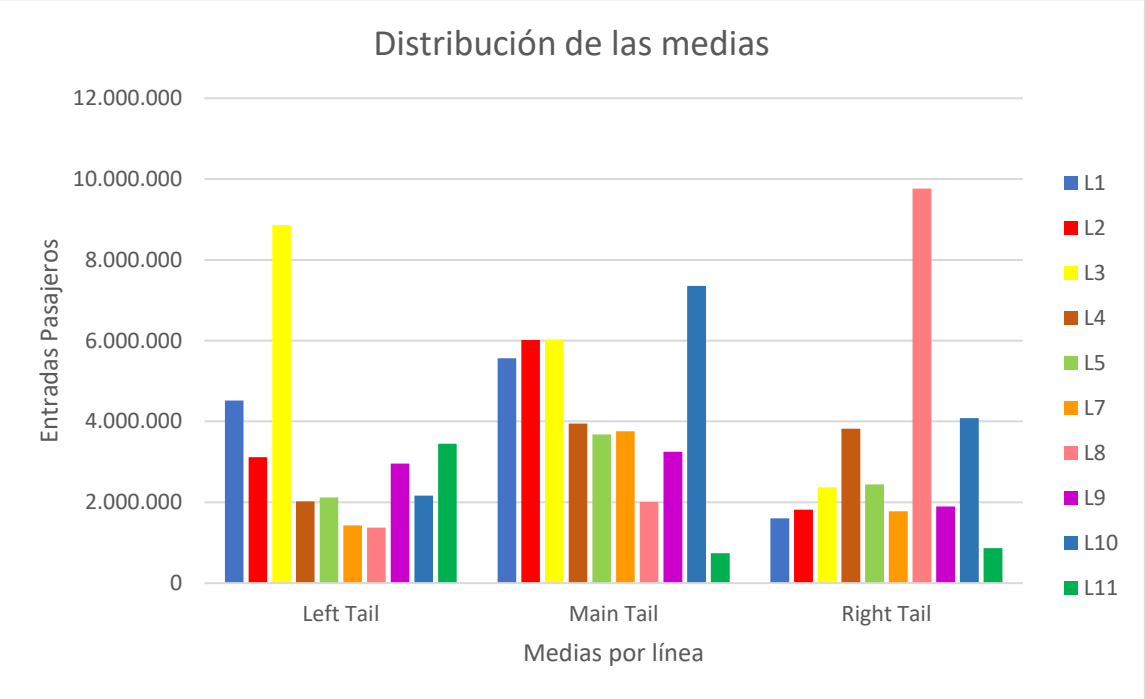
El **Uso de las Estaciones de Metro está explicado en un porcentaje muy alto por la existencia de transbordo en esa estación**, tanto de red de metro como externa, en cualquier de los tres modelos estudiados. Es el factor común para todos los Conjuntos de datos, que son valores esperados.

Lo que podemos extraer observando el segundo Conjunto, es que **el Uso de metro en los barrios de la periferia madrileña viene explicado en un alto porcentaje por la cantidad de Población Joven que habita**, éste es el factor diferencial respecto a los otros dos.

Evaluando el total de las estaciones (Conjunto 1) observamos que los transbordos entre líneas, tanto de metro como de red externa, es el detonante para que haya un mayor uso de esas estaciones por norma general.

En el tercer Conjunto, el del Análisis sobre una nueva incorporación en la línea en los extremos, las variables de transbordo de transporte también tienen un alto grado de importancia; pero **resalta que la Renta per Cápita y la Influencia de la Zona Comercial en la estación tienen papeles importantes**, esto puede deberse a una concentración de renta y de zonas comerciales en el Centro, y los usuarios se mueven de la periferia a la zona central por esta razón.

5.2. Predicciones Extramuestrales



[Figura 19] Distribución media de usuarios según se sitúen en un extremo u otro del trayecto de la línea

En este análisis se proyectará la incorporación de nuevas estaciones en las 3 líneas que mayor demanda de uso tengan en sus extremos:

- En el extremo derecho son la línea 8 final en Nuevos Ministerios, la 10 final en Cuatro Vientos (ya que solo se analiza Madrid Ciudad), y la 4 final en Arguelles.

- En el extremo izquierdo de la distribución, las líneas que mayor media de entradas tienen son la línea 3 final en Moncloa, la línea 1 final en Pinar de Chamartín, y la 11 final en Plaza Elíptica.

El proceso realizado para calcular la estimación de usuarios ha sido el uso del modelo de árboles regresivo explicado en el apartado anterior. Una vez obtenido ese modelo predictivo, el modo de proceder ha sido cambiar el fichero de datos de testeo que teníamos, por un nuevo fichero que incluía aquellas estaciones propuestas para realizar el alargue. El procedimiento realizado para el cálculo de estas variables en las nuevas estaciones ha sido el mismo que el que se llevó a cabo para las demás. Así pues, únicamente ejecutando el algoritmo, el sistema nos ha devuelto el valor normalizado de la estimación de uso, que ha habido que transformar para poder leer e interpretar dichos resultados.

Una duda surgida durante el transcurso de esta sección ha sido la elección del Modelo correcto a utilizar para las estimaciones. Si bien es cierto que el Modelo para el Conjunto 2 tiene un mayor poder predictivo (su ECM es el menor), no es capaz de alcanzar a estaciones que estén en el Centro de Madrid. Así pues, ante esta disyuntiva, he procedido a utilizar el Segundo Modelo para aquellas estaciones que vayan a situarse en la Periferia, y usaré el Tercer Modelo para captar los potenciales usuarios de las estaciones dentro de la Almendra Central.

5.2.1. Caso de Estudio 1: Estimación por el extremo derecho de la distribución.

Si analizamos primero el **extremo derecho de las distribuciones**, vemos que por la línea 4 hay poca posibilidad de alargar la línea y ya lo he realizado para el alargue de la línea 3, explicado más adelante. Y si observamos la línea 10, ya está hecho un alargue hasta Alcorcón que no hemos incorporado al hacer el análisis únicamente para el municipio de Madrid.

Pero es en la línea 8 (Anexo VI [Figura 32]) donde sí se observa una posibilidad de alargar la trayectoria hacia el sur, así pues, las estimaciones de las estaciones propuestas van desde Nuevos Ministerios por el sur atravesando el Centro de Madrid hasta el barrio de Las Águilas:

Órden	Línea 8	Periferia	Modelo usado	Estimación de usuarios	Diferencia con usuarios actuales
1	Atalaya	Si	Modelo 2	1.120.876	Nueva estación
2	Alonso Cano	No	Modelo 3	6.403.563	4.515.157
3	Zurbarán	No	Modelo 3	2.711.908	Nueva estación
4	Colón	No	Modelo 3	6.497.545	5.041.780
5	Recoletos	No	Modelo 3	8.541.486	Nueva estación
6	Banco de España	No	Modelo 3	6.213.749	2.836.551
7	El Prado	No	Modelo 3	2.348.098	Nueva estación
8	Antón Martín	No	Modelo 3	5.962.824	3.278.108
9	Lavapiés	No	Modelo 3	5.985.260	2.378.876
10	Puerta de Toledo	No	Modelo 3	3.129.976	1.502.434
11	Imperial	No	Modelo 3	1.406.757	Nueva estación
12	Ermita del Santo	Si	Modelo 2	1.731.880	Nueva estación
13	San Isidro	Si	Modelo 2	1.692.758	Nueva estación
14	Laguna	Si	Modelo 2	4.070.566	2.090.868
15	Los Yébenes	Si	Modelo 2	2.023.239	Nueva estación
16	Las Águilas	Si	Modelo 2	1.930.747	Nueva estación

[Figura 20] Estimación trayecto línea 8

El sistema nos devuelve la estimación de usuarios. El término Periferia hace referencia a las estaciones situadas fuera de la almendra central (M-30).

5.2.2. Caso de Estudio 2: Estimación por el extremo izquierdo de la distribución

Vamos a analizar las **estimaciones para las estaciones del extremo izquierdo** de los trayectos. En este caso serán las líneas 1, 3 y 11.

Órden	Línea 1	Periferia	Modelo usado	Estimación de usuarios	Diferencia con usuarios actuales
1	Fuente de la Mora	Si	Modelo 2	2.730.268	Nueva estación
2	Alcalá Zamora	Si	Modelo 2	1.000.219	Nueva estación
3	Valdebebas	Si	Modelo 2	1.076.640	Nueva estación

[Figura 21] Estimación trayecto línea 1

Para la línea 1 (Anexo VI [Figura 33]), el alargue se efectúa desde la estación de Pinar de Chamartín hasta una hipotética estación en Valdebebas.

Órden	Línea 3	Periferia	Modelo usado	Estimación de usuarios	Diferencia con usuarios actuales
1	Rosa Luxemburgo	Si	Modelo 2	616.940	Nueva estación
2	Aravaca	Si	Modelo 2	2.265.167	Nueva estación

[Figura 22] Estimación trayecto línea 3

En la línea 3 (Anexo VI [Figura 34]), se efectúa la prolongación desde Moncloa hasta Aravaca, donde se sitúan dos estaciones.

Órden	Línea 11	Periferia	Modelo usado	Estimación de usuarios	Diferencia con usuarios actuales
1	Comillas	Si	Modelo 2	2.234.196	Nueva estación
2	Puente de Praga	No	Modelo 3	2.777.979	Nueva estación
3	Delicias	No	Modelo 3	6.186.290	2.588.248
4	Atocha-Renfe	No	Modelo 3	12.289.100	3.791.007
5	Mariano de Cavia	No	Modelo 3	2.288.391	Nueva estación
6	Conde de Casal	No	Modelo 3	5.178.370	42.941
7	Estrella	Si	Modelo 2	3.239.541	1.321.498
8	Blas de Otero	Si	Modelo 2	1.790.585	Nueva estación
9	La Elipa	Si	Modelo 2	4.477.259	1.738.451
10	Ascao	Si	Modelo 2	4.689.972	2.708.046
11	Ciudad Lineal	Si	Modelo 2	5.148.621	2.399.947
12	El Salvador	Si	Modelo 2	1.232.804	Nueva estación
13	Josefa Valcárcel	Si	Modelo 2	1.435.871	Nueva estación
14	Arturo Soria	Si	Modelo 2	3.391.001	1.395.365
15	Hispanoamérica	No	Modelo 3	1.860.375	Nueva estación
16	Colombia	No	Modelo 3	3.350.472	-522.529
17	Cuzco	No	Modelo 3	3.579.671	67.519
18	Tetuán	No	Modelo 3	3.539.702	178.804
19	Berruguete	No	Modelo 3	2.336.213	Nueva estación
20	Valdezarza	No	Modelo 3	2.139.903	942.719
21	Isla de Oza	No	Modelo 3	1.605.289	Nueva estación
22	Fuentelareyna	Si	Modelo 2	794.427	Nueva estación

[Figura 23] Estimación trayecto línea 11

Para finalizar, la prolongación de la línea 11 (Anexo VI [Figura 32]) va por el Este hacia La Elipa para luego girar por Arturo Soria y acabar en el Oeste en el barrio de Fuentelareyna.

Como se puede observar, hay dos tipos de predicciones: las realizadas para una estación nueva, donde se estima el número anual de usuarios de Metro desde cero; o las predicciones realizadas para una estación existente, pero a la que se añade un valor nuevo en el Transbordo de Metro al tener de este modo correspondencia con otra línea, por ello se estima también la diferencia con la cantidad de usuarios actuales y se analiza la capacidad del modelo para ajustarse a ello.

6. CONCLUSIONES

En este trabajo se analiza el uso de Metro de Madrid que hace la población utilizando determinadas variables relacionadas con el comportamiento de la población, así como la renta y características del metro. La capacidad que tengan las variables independientes para explicar la variabilidad de la dependiente es clave a la hora de sacar

conclusiones tanto a nivel descriptivo de los factores como a nivel predictivo para estimar nuevas estaciones extramuestrales.

La primera de las conclusiones que se extraen de este proyecto trata sobre la importancia de los factores en cada Conjunto. Una primera solución superficial daría como resultado el pensar que el uso de Metro viene condicionado exclusivamente por renta o afluencia de gente en la zona en cuestión; pero después de analizar los resultados, **la conclusión es que el uso que la población hace del Metro es distinto según la zona que se estudie, y en concreto, en las zonas alejadas del centro la cantidad de Población Joven que habite en la zona va a influir de un modo alto la cantidad de usuarios.**

Si ahondamos un poco más en este resultado, vemos que la **importancia de los factores es casi opuesta según la zona que estudiemos. Si antes hemos concluido que la cantidad de Población Joven mantiene relación positiva con el uso de Metro en el segundo Conjunto, la densidad de Población que Habita en el Hogar de las estaciones del Centro (Conjuntos 1 y 2) tiene una relación negativa con la variable dependiente**. Cabe destacar también la importancia que cobran la renta y el grado comercial de la zona en cuestión para estaciones del Centro, en contraste con la Periferia.

Tras examinar los resultados basados en la estimación de nuevas estaciones extramuestrales, se comprueba que **el modelo predictivo es capaz de realizar las estimaciones de forma aceptable para cada nueva estación en discusión (el ECM de los Modelos utilizados para realizar predicciones tiene valor bajo), tanto si se trata de nuevas estaciones como si tiene que ajustarse con nuevas variables, como la de Transbordo de Metro, sobre paradas ya existentes**.

7. LINKS

[1] https://www.alteryx.com/analytics/pages/trial?utm_source=google&utm_medium=cpc&utm_campaign=Southern_Europe_%7C_Brand_%7C_EXACT&utm_source=google&lsm=google&utm_medium=cpc&utm_campaign=Demgen%20Mixed%20-%20Brand_New-UK%20-%20Exact&utm_content=Trial&utm_term=alteryx&utm_adid=257700506543&gclid=CjwKCAjw6djYBRB8EiwAoAF6odLmJM2CokUrRJ55Uj5_Asf_zG7H4QEIPmns4DckOhqsKGSJ_bAjUBoCrU4QAvD_BwE

[2] https://cran.r-project.org/mirrors.html

[3] http://www.experian.es/index.html

[4] http://www-2.munimadrid.es/CSE6/jsps/menuBancoDatos.jsp

[5] http://www.ine.es/

[6] http://www.madrid.org/iestadis/

[7] https://www.metromadrid.es/es/conocenos/infraestructuras/red/

[8] https://www.metromadrid.es/es/viaja_en_metro/red_de_metro/planos/index.html

[9] https://www.disfrutamadrid.com/mapa

[10] https://www.metromadrid.es/es/portal_de_transparencia/inf_econ_presup_estad/datos_estadisticos/index.html

[11] . https://www.ucm.es/fundamentos-analisis-economico2/material-preparado-por-los-profesores-de-la-asignatura

[12] http://www2.uned.es/fac-poli/netscp/comollegar/n2a5.htm

[13] https://www.quora.com/What-is-the-simple-explanation-of-M5P-M5-model-trees-algorithm-in-Machine-learning-Data-Mining

8. BIBLIOGRAFÍA

Anibarro García, Javier; INECO (2016). “La competitividad del transporte en España”, Observatorio del Transporte y la Logística en España.

Ayuntamiento de Madrid (2014). “Plan de Movilidad Urbana Sostenible (PMUS) de la ciudad de Madrid”, Ayuntamiento de Madrid.

Daniel-Cardozo, Osvaldo; García Palomares, Juan Carlos y Gutiérrez Puebla, Javier (2008). “Modelos de demanda potencial de viajeros en redes de transporte público:

aplicaciones en el Metro de Madrid", III Seminario Internacional de Ordenamiento territorial – la Interdisciplina en el Ordenamiento Territorial. Universidad Complutense de Madrid y Universidad Nacional del Nordeste (Resistencia, Argentina).

Daniel-Cardozo, Osvaldo; García Palomares, Juan Carlos y Gutiérrez Puebla, Javier (2010). "Influencia de la Morfología Urbana en la demanda de transporte público: análisis mediante SIG y modelos de regresión múltiple", GeoFocus.

Daniel-Cardozo, Osvaldo; García Palomares, Juan Carlos y Gutiérrez Puebla, Javier (2012). "Regresión Geográficamente Ponderada (GWR) y estimación de la demanda de las estaciones del Metro de Madrid", XV Congreso Nacional de Tecnologías de la Información Geográfica, Madrid; Universidad Complutense de Madrid y Universidad Nacional del Nordeste (Resistencia, Argentina).

Cristóbal Pinto, Carlos; Gómez Cerdá, Gabriel y Gutiérrez Puebla, Javier (2000). "Accesibilidad peatonal a la red de metro de Madrid: efectos del Plan de Ampliación 1995-99", Anales de Geografía de la Universidad Complutense.

Gromping, Ulrike (2006). "Relative Importance for Linear Regression in R: The Package relaimpo", Journal of Statistical Software. Librería de R.

Guisande González, Cástor y Vaamonde Liste, Antonio (2012). "Gráficos Estadísticos y Mapas con R, Ediciones Díaz de Santos.

Hofner, Benjamin (2017). "PapeR". Librería de R.

Kuhn, Max (2018). "Caret: Classification and Regression Training". Librería de R

Lanz, Brett (2015) "Machine Learning with R, Second edition", PACKT Publishing.

Méndez Suárez, Mariano (2018). "Análisis de Datos con R. Una Aplicación a la Investigación de Mercados", ESIC.

Peters, Andrea y Hothorn, Torsten (2017). "ipred: Improved Predictors". Librería de R.

Revelle, William (2018). "Psych: Procedures for Psychological, Psychometric, and Personality Research", Northwestern University. Librería de R.

Servicio Planificación y Estudios de Operación del Ayuntamiento de Madrid (2018). "Evolución de la Demanda, Cierre de año 2017". Ayuntamiento de Madrid.

Sotoca, Sonia (2012). "Transparencias de Econometría". Mimeo UCM.

Wei, Taiyun and Simko, Viliam (2017). "corrplot: Visualization of a Correlation Matrix". Librería de R.

Wickham, Hadley (2009). "ggplot2: Elegant Graphics for Data Analysis", Springer-Verlag New York. Librería de R.

Wickham, Hadley (2017). "scales: Scale Functions for Visualization". Librería de R.

Wickham, Hadley; Francois, Romain; Henry, Lionel y Müller, Kirill (2017). "A Grammar of Data Manipulation". Librería de R.

Xie, Yihui. “Tinytex: Helper Functions to Install and Maintain 'TeX Live', and Compile 'LaTeX' Documents”. Librería de R.

Zeileis, Achim y Hothorn, Torsten (2002). “Diagnostic Checking in Regression Relationships”, R News. Librería de R.

Zeileis, Achim (2004). “Econometric Computing with HC and HAC, Covariance Matrix Estimators”, Journal of Statistical Software. Librería de R.

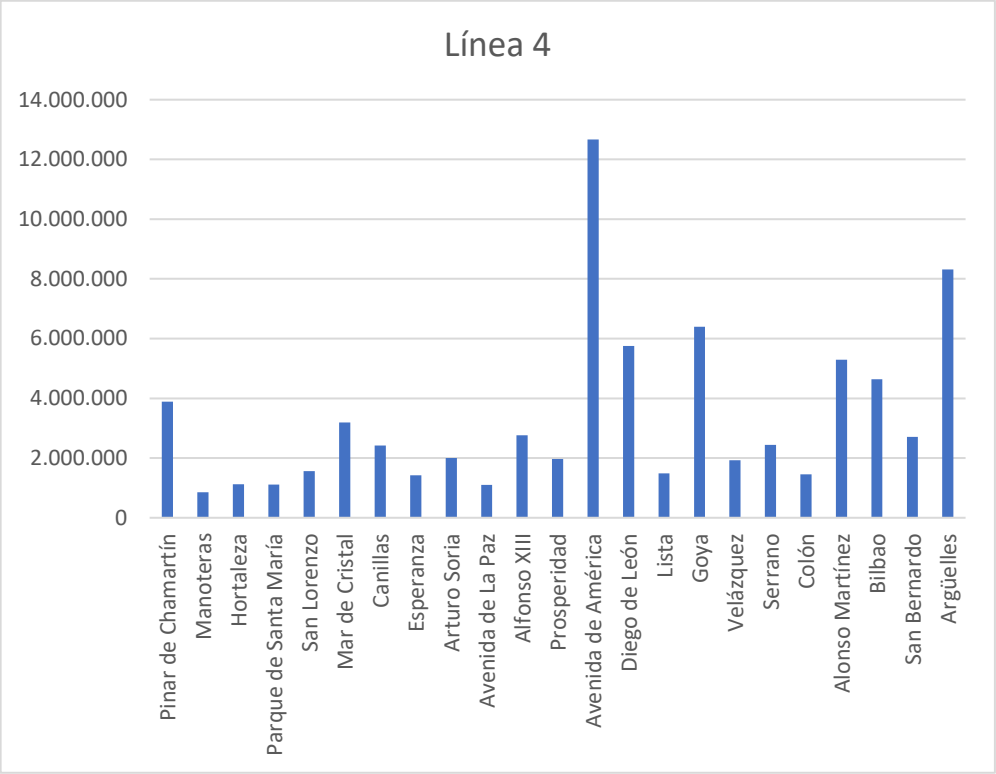
9. ANEXOS

I. TABLA DE VARIABLES UTILIZADAS EN EL PROYECTO

I1.Uso_estacion	Score Uso (Entradas) Metro frente a total
I2.RentaPC	Score Renta disponible de la población del centroide (2014)
I3.Transbordo_metro	Transbordo con red de metro (Dummy)
I4.Transbordo_Externo	Transbordo con red de transporte externa: Cercanías, metro ligero, Intercambiador de autobuses (Score)
I5.Estacion_adaptada	Estación adaptada a minusválidos (Dummy)
I6.Estacion_parking	Estación con párking (Dummy)
I7.Influencia_Comercial	Zona influencia turístico-cultural-comercial (Dummy)
I8.Poblacion_Joven	Score Población Joven (10-24) del centroide
I9.Poblacion_Adulta	Score Población Adulta (25-65) del centroide
I10.Poblacion_Senior	Score Población Senior (66-85) del centroide
I11.Zona_en_auge	Ratio de zona en potencial auge poblacional ((Pjov+Padu)/(Ptot)) en el centroide
I12.Poblacion_edad_trabajar	Población potencial en edad de trabajar (Padu/Ptot) del centroide
I13.Zona_Envejecida	Ratio de envejecimiento (Psenior/Pjoven) de la zona del centroide
I14.DensidadPobl_Hogar	Relación entre número de habitantes y número de hogares (NºHab/NºHog) que hay en el centroide
I15.Estaciones_Hasta_Transbordo	Estaciones restantes hasta transbordo
I16.Numero_hogares	Score de número de hogares en la sección censal

I17.Renta_media_hogar	Score Renta media por Hogar en el centroide
I18.GeneroPoblacion	Relación entre el número de mujeres y hombres (NºMuj/NºHom) que habitan en el centroide
I19.Tasa_Inmigracion	Relación entre la población Nacional y extranjera (NºNac/NºExtr) que habita en el centroide

II. DISTRIBUCIÓN NO NORMAL



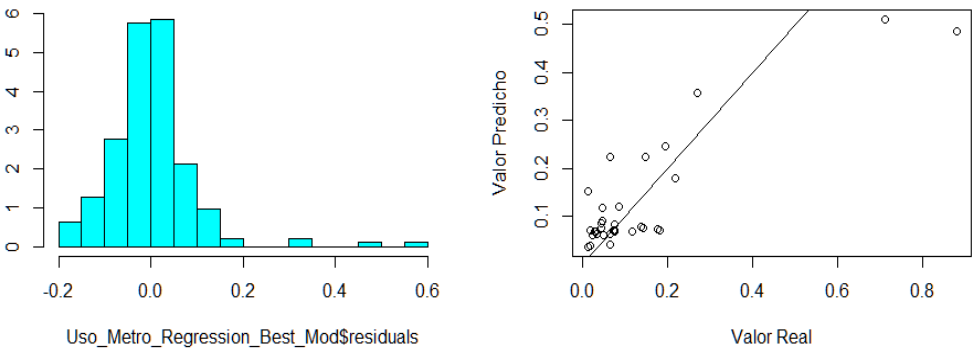
III. ANEXO CONJUNTO 1

Tabla del análisis de la varianza (ANOVA)

Model 1	I1.Uso_estacion~I2.RentaPC+I3.Transbordo_metro+ I4.Transbordo_Externo+ I5.Estacion_adaptada + I6.Estacion_parking +I7.Influencia_Comercial+I8.Poblacion_Joven+ I12.Poblacion_edad_trabajar +I13.Zona_Envejecida+I14.DensidadPobl_Hogar+ I15.Estaciones_Hasta_Transbordo+I18.GeneroPoblacion+ I19.Tasa_Inmigracion
Model 2	I1.Uso_estacion~I2.RentaPC+I3.Transbordo_metro+ I4.Transbordo_Externo+I5.Estacion_adaptada+ I14.DensidadPobl_Hogar + I19.Tasa_Inmigracion

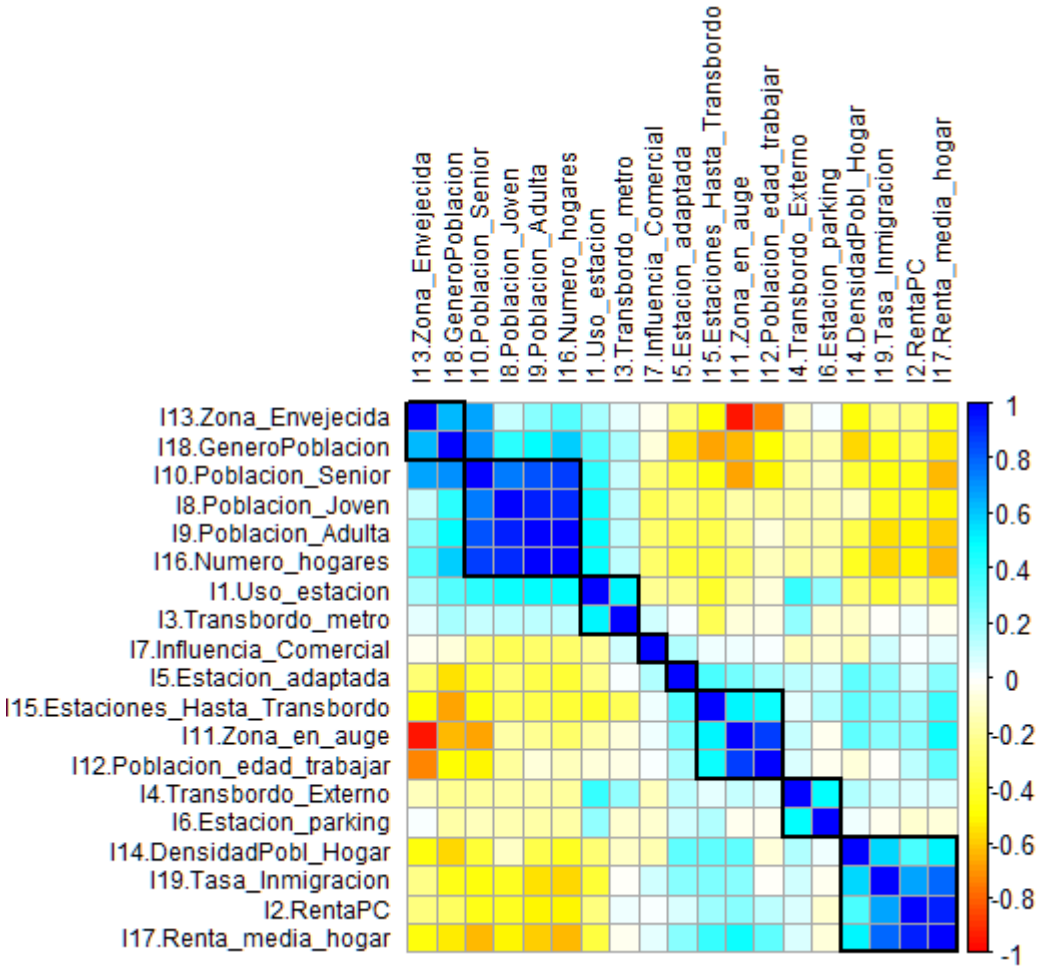
MODEL	RES. DegrFree	RSS	DegrFree	SUM OF SQ	F	Pr(>F)
1	174	1.5513	0	0	0	0
2	181	1.5818	-7	-0.030576	0.4899	0.8411

[Figura 24] ANOVA del Conjunto 1



[Figura 25] Histograma de Residuos y ECM del Conjunto 1

IV. ANEXO CONJUNTO 2



[Figura 26] Matriz Correlaciones Conjunto 2

Análisis de Criterio de Akaike

REGRESSION	DEGREE OF FREEDOM	CRITERIO AKAIKE
Uso_Metro_Regression_05	15	-128.2752
Uso_Metro_Regression_Best_Mod	11	-135.6171

Tabla del análisis de la varianza (ANOVA)

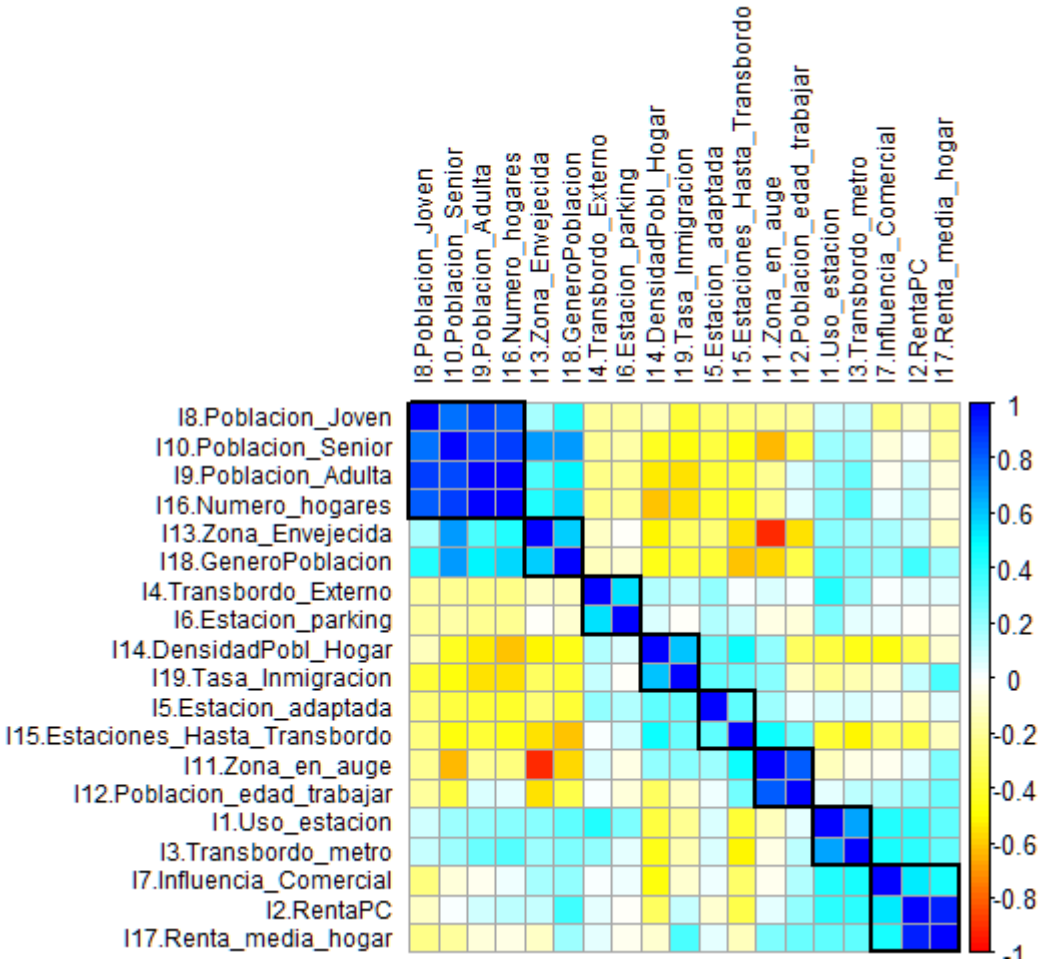
Model 1	I1.Uso_estacion~I2.RentaPC+I3.Transbordo_metro+ I4.Transbordo_Externo+ I5.Estacion_adaptada + I6.Estacion_parking +I7.Influencia_Comercial+I8.Poblacion_Joven+ I12.Poblacion_edad_trabajar +I13.Zona_Envejecida+I14.DensidadPobl_Hogar+ I15.Estaciones_Hasta_Transbordo+I18.GeneroPoblacion+ I19.Tasa_Inmigracion
Model 2	I1.Uso_estacion~I3.Transbordo_metro+ I4.Transbordo_Externo+I5.Estacion_adaptada+I6.Estacion_parking +I8.Poblacion_Joven+I12.Poblacion_edad_trabajar +I13.Zona_Envejecida+I15.Estaciones_Hasta_Transbordo+ I19.Tasa_Inmigracion

MODEL	RES. DegrFree	RSS	DegrFree	SUM OF SQ	F	Pr(>F)
1	85	1.1718	0	0	0	0
2	89	1.1796	-4	-0.0078148	0.1417	0.9662

[Figura 27] Criterio de información de Akaike y Contraste Hipótesis ANOVA Conjunto 2

Modelo de decisión Conjunto 2

V. ANEXO CONJUNTO 3



[Figura 28] Matriz de correlaciones Conjunto 3

Análisis de Criterio de Akaike

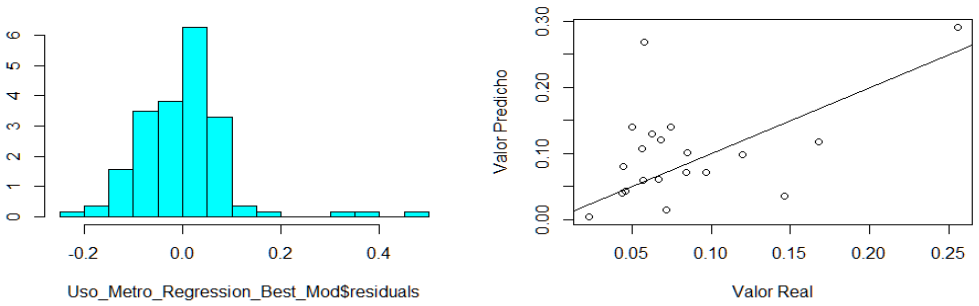
REGRESSION	DEGREE OF FREEDOM	CRITERIO AKAIKE
Uso_Metro_Regression_05	15	-185.0814
Uso_Metro_Regression_Best_Mod	10	-194.6074

Tabla del análisis de la varianza (ANOVA)

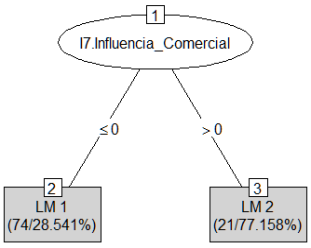
Model 1	I1.Uso_estacion~I2.RentaPC+I3.Transbordo_metro+ I4.Transbordo_Externo+ I5.Estacion_adaptada + I6.Estacion_parking +I7.Influencia_Comercial+I8.Poblacion_Joven+ I12.Poblacion_edad_trabajar +I13.Zona_Envejecida+I14.DensidadPobl_Hogar+ I15.Estaciones_Hasta_Transbordo+I18.GeneroPoblacion+ I19.Tasa_Inmigracion
Model 2	I1.Uso_estacion~I2.RentaPC+I3.Transbordo_metro+ I4.Transbordo_Externo+I5.Estacion_adaptada+I7.Influencia_Comercial +I8.Poblacion_Joven +I13.Zona_Envejecida + I19.Tasa_Inmigracion

MODEL	RES. DegrFree	RSS	DegrFree	SUM OF SQ	F	Pr(>F)
1	101	1.0375	0	0	0	0
2	106	1.0417	-5	-0.0042853	0.0834	0.9947

[Figura 29] Criterio de Akaike y Test ANOVA Conjunto 3

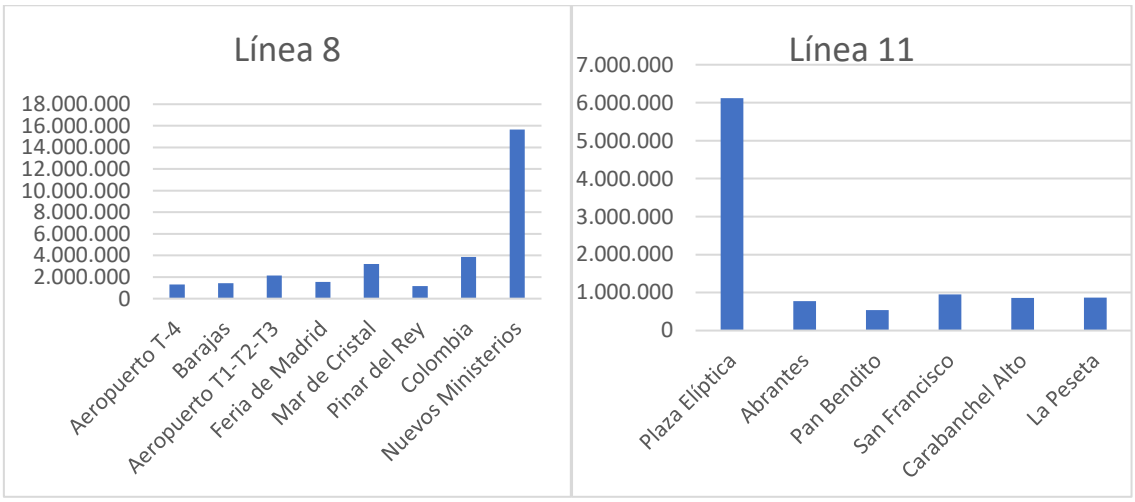


[Figura 30] Histograma de Residuos y ECM del Conjunto 3

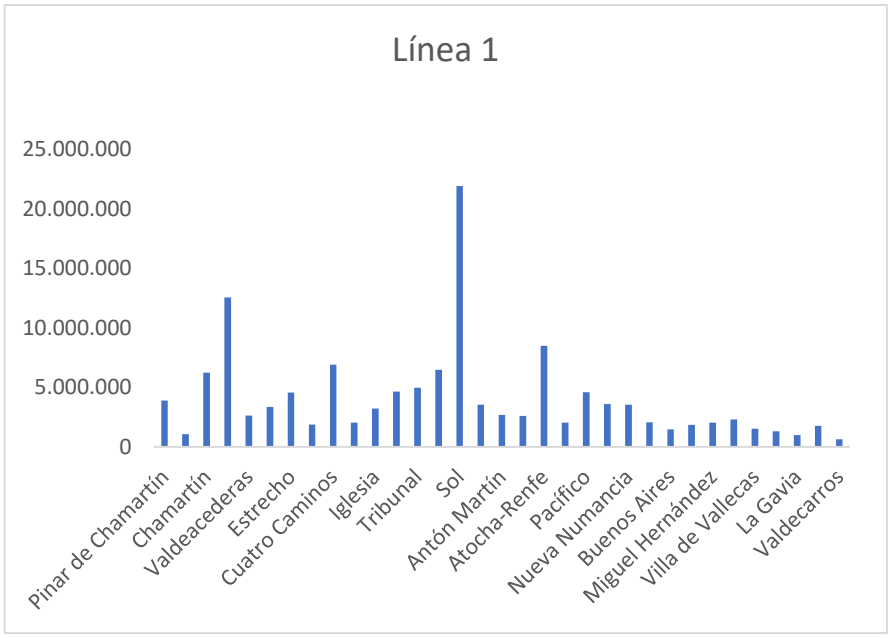


[Figura 31] Árbol de decisión Conjunto 3

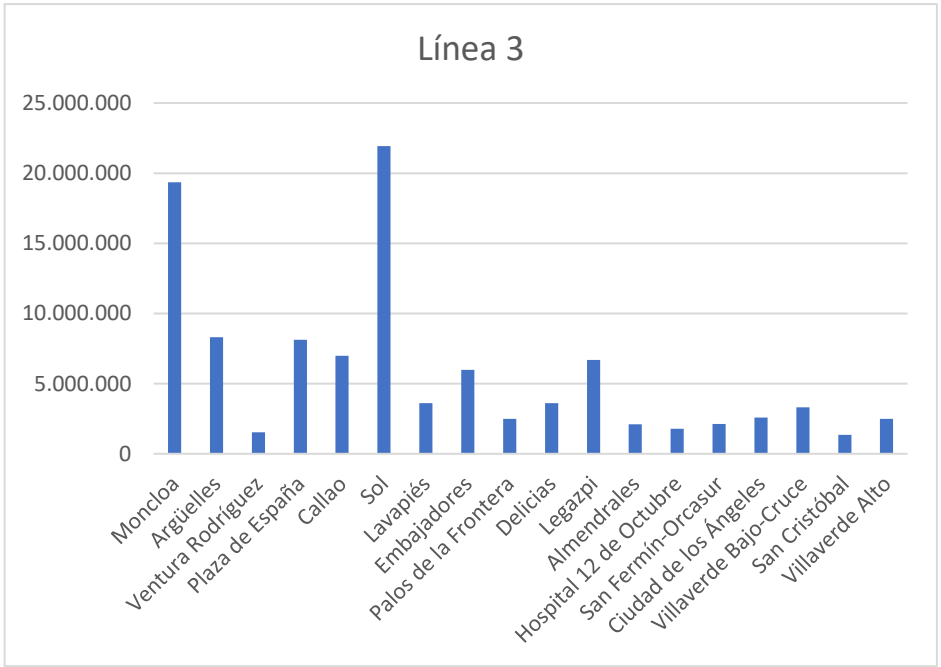
VI. ANEXO PREDICCIÓN DE USO EN NUEVAS ESTACIONES



[Figura 32] Distribución de usuarios en Líneas 8 y 11



[Figura 33] Distribución usuarios línea 1



[Figura 34] Distribución usuarios línea 3

VII. ANEXO ALTERYX